Perceptually Weighted Analysis-by-Synthesis Vector Quantization for Low Bit Rate MFCC Codec

Gang Min, Xiongwei Zhang, Xia Zou, and Jibin Yang

***Abstract*—This letter presents a perceptually weighted analysis -by-synthesis vector quantization (VQ) algorithm for low bit rate MFCC codec. Different from conventional VQ of MFCCs vector, this algorithm uses an analysis-by-synthesis technique and aims to minimize the perceptually weighted spectral reconstruction distortion rather than the distortion of MFCCs vector itself. Also, to reduce the computational complexity, we propose a practical suboptimal codebook searching technique and embed it into the split and multistage vector quantization framework. Objective and subjective experimental results for Mandarin speech show that the proposed algorithm yields intelligible and natural sounding speech for speech coding at 600–2400 bit/s. Compared to current VQ in MFCC codec, the output speech quality is substantially improved in terms of frequency-weighted segmental SNR, STOI, PESQ and MOS score.**

***Index Terms*—MFCCs, vector quantization, speech coding, analysis-by-synthesis.**

I. Introduction

MFCC codec attempts to encode the speech signal through quantization of mel-frequency cepstral coefficients (MFCCs), which provides a promising new approach for speech coding throughout 600–4800 bit/s. The speech quality of MFCC codec even exceeds the output of state-of-the-art MELPe codec [1]–[2]. Also, high-resolution MFCCs vector encoded in MFCC codec could be easily down-converted to low-resolution MFCCs vector for distributed speech recognition (DSR) in ETSI Aurora DSR standard [3]. Despite natural sounding speech yielded from MFCC codec, there is still room for further improving the speech quality. For example, there exists the phenomenon of spectrum smearing since the triangle frequency window used for MFCCs extraction is overlapped. Moreover, the quantization process of MFCCs vector will further aggravate this problem, which degrades the articulation of the coded speech for MFCC codec.

Vector quantization (VQ) plays important role in reducing the bit rate of speech coding. However, the quantization error increases rapidly with the decreasing of quantization bits [4], which is the main reason accounting for degradation of speech quality. In the current MFCC codec, MFCCs vector is directly quantized with the objective of minimizing the quantization distortion of itself, i.e., the codeword of minimum square error is selected as the quantized vector [2]. Yet, this conventional VQ method can not straightly illustrate the effect of quantization on the final speech distortion. Inspired by the perceptual weighting technique used in the low bit rate codec [5]–[6], which considers the auditory masking properties of the human's ear, we change the objective of VQ of MFCCs vector as: *the codeword of minimum perceptually weighted spectral reconstruction distortion is selected as the quantized MFCCs vector.* To achieve this objective, we propose a new framework for VQ of MFCCs vector to minimize the end-to-end perceptually weighted spectral distortion for speech signal rather than the quantization distortion for MFCCs vector. The proposed method strategically use a closed-loop technique known as *analysis-by-synthesis* (AbS), which has been broadly used for speech coding and the quantization of compressed sensing measurements [7]–[10]. The synthesis step employs a speech power spectrum reconstruction technique for measuring the effect of MFCCs vector quantization on the final speech quality, and the analysis step is performed followed by the synthesis step in order to select an appropriate codeword to minimize the perceptually weighted spectral distortion. To the best of our knowledge, the perceptually weighted AbS approach has not been used for VQ of MFCCs vector earlier, which is shown to provide a much better speech quality than the conventional VQ method that directly quantizes the MFCCs vector itself in an open-loop fashion. Since AbS requires higher computational cost, a low complexity suboptimal codebook searching technique is also proposed.

Notations: for the k^{th} speech frame $(k = 1, 2, ..., K)$, $\boldsymbol{y}_k$ and $\boldsymbol{f}_k$ denote the power spectrum and MFCCs vectors, $\hat{\boldsymbol{y}}_k$ and $\hat{\boldsymbol{f}}_k$ denote their quantized values, $\boldsymbol{W}_k$ denotes the perceptual weighting matrix. $\mathcal{F} = \{\tilde{\boldsymbol{f}}_1, \tilde{\boldsymbol{f}}_2, ..., \tilde{\boldsymbol{f}}_J\}$, $\mathcal{F}_s = \{\tilde{\boldsymbol{f}}_1, \tilde{\boldsymbol{f}}_2, ..., \tilde{\boldsymbol{f}}_I\}$ denote the codebook and suboptimal codebook of MFCCs vector, respectively. $\tilde{\boldsymbol{f}}_j, j = 1, 2, ..., J$ denotes the codeword (also called centroid in this paper). $\tilde{\boldsymbol{y}}_j$ denotes the power spectrum reconstructed from $\tilde{\boldsymbol{f}}_j$. $\mathcal{C}_j, j = 1, 2, ...J$ denotes the j^{th} cluster in the codebook training phase.

II. Analysis-by-Synthesis Vector Quantization

A. Mel-frequency cepstral coefficients

The extracting procedure of MFCCs begins with enframing the speech waveforms $x(n)$ by a window $w(n)$,

$$x_m(n) = x(mR + n)\, w(n) \tag{1}$$

where $N(0 \le n \le N-1)$ is the window length, R is the frame shift, m is the frame index. Then, the speech frame could be concisely denoted as,

$$\boldsymbol{x} = [x_m(0), x_m(1), ..., x_m(N-1)]^{\mathsf{T}} \tag{2}$$

The power spectrum of each speech frame is,

$$\boldsymbol{y} = |\mathrm{F}\{\boldsymbol{x}\}|^2 \tag{3}$$

G. Min is with the Lab. of Intelligent information processing, PLA University of Science and Technology, Nanjing, 210007 China. He is also with Xi'an communications Institute, Xi'an, 710106 China. e-mail: mgxaty@gmail.com.

Manuscript uploaded June 18, 2016.

where $\mathrm{F}\{\boldsymbol{x}\}$ is the N-point FFT of $\boldsymbol{x}$, $|\cdot|$ denotes the modulus of a complex number.

The latter $N/2-1$ elements of $\boldsymbol{y}$ will be discarded due to the symmetry. Then, the power spectrum is Mel-filtered by a set of weighting functions, i.e., the Mel-scale weighting matrix $\boldsymbol{\Phi} \in \mathbf{R}^{M\times(N/2+1)}$, where M is the number of Mel-filter bands. Generally, $\boldsymbol{\Phi}$ is designed based on human perception of pitch frequency and implemented in the form of a bank of filters, each filter is with a triangular frequency response [2]. Finally, MFCCs vector is computed through the $\log(\cdot)$ and discrete cosine transform (DCT),

$$\boldsymbol{f} = \mathrm{DCT}\{\log(\boldsymbol{\Phi y})\} = \mathbf{D}\log(\boldsymbol{\Phi y}) \tag{4}$$

where $\mathbf{D}$ is the $M \times M$ DCT matrix.

The power spectrum $\boldsymbol{y}$ could be approximately reconstructed from MFCCs vector $\boldsymbol{f}$ as follows [1]–[2],

$$\boldsymbol{y} = \boldsymbol{\Phi}^{\dagger}\exp(\mathbf{D}^{-1}\boldsymbol{f}) = (\boldsymbol{\Phi}^{\mathsf{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathsf{T}}\exp(\mathbf{D}^{-1}\boldsymbol{f}) \tag{5}$$

where $\boldsymbol{\Phi}^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of $\boldsymbol{\Phi}$.

B. Distance measure for VQ of MFCCs vector

Conventionally, the MFCCs vector $\boldsymbol{f}$ is directly quantized using the Euclidean distance measure [2],

$$d\left(\boldsymbol{f},\hat{\boldsymbol{f}}\right) = \left\|\boldsymbol{f}-\hat{\boldsymbol{f}}\right\|_2^2 = \sum_{i=0}^{M-1}\left(f_i-\hat{f}_i\right)^2 \tag{6}$$

here, we consider the perceptual weighting filter $P(z)$,

$$P(z) = \frac{H\left(\gamma^{-1}z\right)}{H\left(\beta^{-1}z\right)} = \frac{1-\sum_{i=1}^{p}a_i\beta^i z^{-i}}{1-\sum_{i=1}^{p}a_i\gamma^i z^{-i}} \tag{7}$$

where $H(z) = 1\Big/\left(1-\sum_{i=1}^{p}a_i z^{-i}\right)$ is the linear prediction (LP) synthesis filter, a_i is the short-term LP coefficients and p is the prediction order. $0<\gamma<\beta\leq 1$ are the perceptual weighting factors that control the energy of the error in the formant regions [11]. Referring to the EVRC codec, we choose $\beta = 0.9, \gamma = 0.5$, respectively [12].

Let $P(\omega)$ denote the frequency response of perceptual weighting filter defined in (7),

$$P(\omega) = P(z)\,|_{z=e^{j\omega}} \tag{8}$$

In the spectral domain, the perceptual weighting matrix $\boldsymbol{W}$ could be expressed as a diagonal matrix,

$$\boldsymbol{W} = \begin{bmatrix} w_0 & 0 & \cdots & 0 \\ 0 & w_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & w_{N/2} \end{bmatrix} \tag{9}$$

here,

$$w_i = P(\omega)\Big|_{\omega=\frac{2\pi}{N}i}, \quad i = 0,1,...,N/2. \tag{10}$$

As mentioned above, N is the FFT length.

Consequently, the perceptually weighted spectral distortion between the original power spectrum $\boldsymbol{y}$ and the quantized power spectrum $\hat{\boldsymbol{y}}$ could be expressed as,

$$d(\boldsymbol{y},\hat{\boldsymbol{y}}) = \sum_{i=0}^{N/2} w_i(y_i-\hat{y}_i)^2 = (\boldsymbol{y}-\hat{\boldsymbol{y}})^{\mathsf{T}}\boldsymbol{W}(\boldsymbol{y}-\hat{\boldsymbol{y}}) \tag{11}$$

C. Overview of perceptually weighted AbS VQ

As is shown in Fig. 1, the proposed perceptually weighted AbS VQ mainly consists of two steps: a synthesis step that reconstructs the speech power spectrum from the MFCCs codeword $\tilde{\boldsymbol{f}}_j$ and an analysis step that extracts the power spectrum of the k^{th} speech frame and calculates the perceptually weighted spectral distortion between $\boldsymbol{y}_k$ and $\tilde{\boldsymbol{y}}_j$. These two steps will be repeated J times for searching the whole codebook $\mathcal{F}$. Finally, the codeword of minimum $d(\boldsymbol{y}_k,\tilde{\boldsymbol{y}}_j)$ is selected as the quantized MFCCs vector.

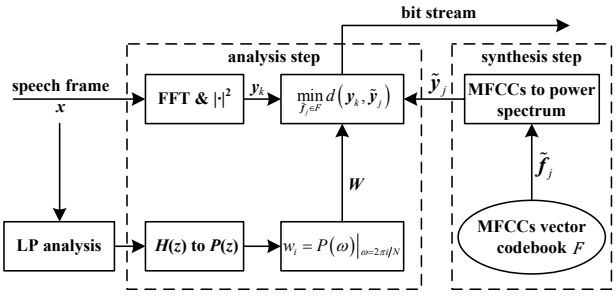

Fig. 1. Diagram of perceptually weighted AbS VQ.

D. Codebook training

To train the codebook $\mathcal{F}$, there are two problems to resolve. The first is how to assign each training sample into a cluster, the second is how to determine the centroid of each cluster. Conventionally, the first can be resolved via the nearest neighbor criterion. However, the second needs further derivations.

Let $\tilde{\boldsymbol{y}}_j$ denote the power spectrum reconstructed from the centroid of the j^{th} cluster. For any training sample $\boldsymbol{y}_k$,

$$d(\boldsymbol{y}_k,\tilde{\boldsymbol{y}}_j) = d\left(\boldsymbol{y}_k,\boldsymbol{\Phi}^{\dagger}\exp(\mathbf{D}^{-1}\tilde{\boldsymbol{f}}_j)\right) \tag{12}$$

Then, $\boldsymbol{y}_k$ can be assigned into a cluster $\mathcal{C}_j$ according to the nearest neighbor criterion,

$$\begin{aligned}\mathcal{C}_j = \{\boldsymbol{y}_k \,|\, & d(\boldsymbol{y}_k,\tilde{\boldsymbol{y}}_j) < d(\boldsymbol{y}_k,\tilde{\boldsymbol{y}}_i), k = 1,2,...,K, \\ & i = 1,2,...,J, i\neq j\}\end{aligned} \tag{13}$$

Consequently, for K speech frames as training, the total distortion is,

$$\begin{aligned} e_t &= \sum_{k=1}^{K} d(\boldsymbol{y}_k,\hat{\boldsymbol{y}}_k) = \sum_{j=1}^{J}\sum_{\boldsymbol{y}_k\in\mathcal{C}_j} d(\boldsymbol{y}_k,\tilde{\boldsymbol{y}}_j) \\ &= \sum_{j=1}^{J}\sum_{\boldsymbol{y}_k\in\mathcal{C}_j}(\boldsymbol{y}_k-\tilde{\boldsymbol{y}}_j)^{\mathsf{T}}\boldsymbol{W}_k(\boldsymbol{y}_k-\tilde{\boldsymbol{y}}_j) \\ &= \sum_{j=1}^{J}\sum_{\boldsymbol{y}_k\in\mathcal{C}_j}\left\|\sqrt{\boldsymbol{W}_k}(\boldsymbol{y}_k-\tilde{\boldsymbol{y}}_j)\right\|_2^2 \\ &= \sum_{j=1}^{J}\sum_{\boldsymbol{y}_k\in\mathcal{C}_j}\left\|\sqrt{\boldsymbol{W}_k}\left(\boldsymbol{y}_k-\boldsymbol{\Phi}^{\dagger}\exp(\mathbf{D}^{-1}\tilde{\boldsymbol{f}}_j)\right)\right\|_2^2 \end{aligned} \tag{14}$$

let $e_j = \sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \left\| \sqrt{\boldsymbol{W}_k} \left(\boldsymbol{y}_k - \boldsymbol{\Phi}^{\dagger} \exp\left(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j \right) \right) \right\|_2^2$ denote the total distortion associated with all samples which belong to the j^{th} cluster, then e_t could be concisely represented as,

$$e_t = \sum_{j=1}^{J} e_j \tag{15}$$

With respect to the centroid $\tilde{\boldsymbol{f}}_j$, e_t could be minimized,

$$\frac{\partial e_t}{\partial \tilde{\boldsymbol{f}}_j} = \frac{\partial \sum_{j=1}^{J} e_j}{\partial \tilde{\boldsymbol{f}}_j} = \frac{\partial e_j}{\partial \tilde{\boldsymbol{f}}_j} \tag{16}$$

Applying the chain rule and setting $\frac{\partial e_t}{\partial \tilde{\boldsymbol{f}}_j} = 0$,we have,

$$\frac{\partial \boldsymbol{\Phi}^{\dagger} \exp\left(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j \right)}{\partial \tilde{\boldsymbol{f}}_j^{\mathrm{T}}} \frac{\partial e_j}{\partial \boldsymbol{\Phi}^{\dagger} \exp\left(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j \right)} = 0 \tag{17}$$

Let us recall the definition of e_j, it is apparent that the solution of (18) is also the solution of (17),

$$\frac{\partial \sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \left\| \sqrt{\boldsymbol{W}_k} \left(\boldsymbol{y}_k - \boldsymbol{\Phi}^{\dagger} \exp(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j) \right) \right\|_2^2}{\partial \boldsymbol{\Phi}^{\dagger} \exp(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j)} = 0 \tag{18}$$

that is,

$$\left(\sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k \right) \boldsymbol{\Phi}^{\dagger} \exp\left(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j \right) = \sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k \boldsymbol{y}_k \tag{19}$$

$\boldsymbol{W_k}$ is a diagonal non-zero matrix, then $\sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k$ is non-singular. Hence, the optimal centroid for the j^{th} cluster is,

$$\tilde{\boldsymbol{f}}_j = \mathbf{D} \log \left(\boldsymbol{\Phi} \left(\sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k \right)^{-1} \sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k \boldsymbol{y}_k \right) \tag{20}$$

In summary, the codebook training algorithm for perceptually weighted AbS VQ is shown in **Algorithm1**.

E. Low complexity suboptimal codebook searching

Different from the codebook training phase, the codebook searching procedure should be performed in an online fashion, so the computational complexity becomes a major issue. In the primary perceptually weighted AbS VQ scheme in Fig.1, we should reconstruct speech power spectrum from each codeword $\tilde{\boldsymbol{f}}_j$ in $\mathcal{F}$ and calculate the corresponding perceptually weighted spectral distortion. The computational complexity is too high because the number of codewords in $\mathcal{F}$, i.e., J is very large. Consequently, we propose a low complexity suboptimal codebook searching technique as is shown in Fig.2. Through the conventional VQ of MFCCs vector $\boldsymbol{f}_k$, K optimal candidate codewords are selected from $\mathcal{F}$ to constitute the suboptimal codebook $\mathcal{F}_s$. For $\forall \tilde{\boldsymbol{f}}_j \in \mathcal{F}_s, \forall \tilde{\boldsymbol{f}}_i \in \mathcal{F} \backslash \mathcal{F}_s$,

$$\left\| \boldsymbol{f}_k - \tilde{\boldsymbol{f}}_j \right\|_2^2 < \left\| \boldsymbol{f}_k - \tilde{\boldsymbol{f}}_i \right\|_2^2 \tag{21}$$

Only the codewords in $\mathcal{F}_s$ are used for the reconstruction of speech power spectrum and the calculation of perceptually weighted spectral distortion, so the computational complexity will be reduced dramatically, since we will choose $I \ll J$.

Algorithm 1 Codebook training algorithm.

Input: training samples $\mathcal{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_K\}$, weighting matrixes $\mathcal{W} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, ..., \boldsymbol{W}_K\}$
Output: MFCCs codebook $\mathcal{F} = \{\tilde{\boldsymbol{f}}_1, \tilde{\boldsymbol{f}}_2, ..., \tilde{\boldsymbol{f}}_J\}$
1: **Initialization**: $n = 1, e_t^{(0)} = 0$, T = 50, $\delta = 0.1$, $\mathcal{F}$ is initialized randomly.
2: **while** $n \leq \mathrm{T}, \triangle e_t \geq \delta$ **do**
3: // Line 4 reconstructs the power spectrum $\tilde{\boldsymbol{y}}_j$ from $\tilde{\boldsymbol{f}}_j$:
4: $\tilde{\boldsymbol{y}}_j = \boldsymbol{\Phi}^{\dagger} \exp\left(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j \right)$
5: // Line 6 assigns each sample $\boldsymbol{y}_k$ into a cluster $\mathcal{C}_j$:
6: $\mathcal{C}_j = \{ \boldsymbol{y}_k \,|d\left(\boldsymbol{y}_k, \tilde{\boldsymbol{y}}_j\right) < d\left(\boldsymbol{y}_k, \tilde{\boldsymbol{y}}_i\right), i = 1, 2, ..., J, i \neq j \}$
7: // Line 8 updates the cluster centroid $\tilde{\boldsymbol{f}}_j$:
8: $\tilde{\boldsymbol{f}}_j = \mathbf{D} \log \left(\boldsymbol{\Phi} \left(\sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k \right)^{-1} \sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \boldsymbol{W}_k \boldsymbol{y}_k \right)$
9: // Line 10 computes the total distortion e_t:
10: $e_t^{(n)} = \sum_{j=1}^{J} \sum_{\boldsymbol{y}_k \in \mathcal{C}_j} \left\| \sqrt{\boldsymbol{W}_k} \left(\boldsymbol{y}_k - \boldsymbol{\Phi}^{\dagger} \exp\left(\mathbf{D}^{-1} \tilde{\boldsymbol{f}}_j \right) \right) \right\|_2^2$
11: // Line 12 computes the variation of e_t, $\triangle e_t$:
12: $\triangle e_t = \| e_t^{(n)} - e_t^{(n-1)} \|_2^2$
13: $n = n + 1$
14: **end while**

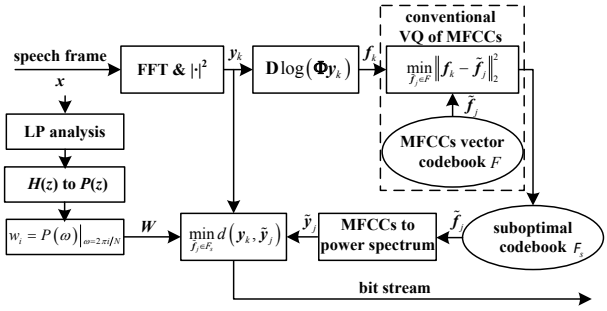

Fig. 2. Diagram of low complexity suboptimal codebook searching.

F. AbS based split and multistage VQ

Furthermore, from the perspective of practical applications, split VQ (SVQ) or multistage VQ (MSVQ) is usually adopted as an alternative of direct VQ to reduce the storage and computational complexity. In order to embed AbS VQ into the SVQ and MSVQ framework, we will keep Q optimal candidate codewords in each sub-vector codebook or sub-stage codebook to constitute the final suboptimal codebook $\mathcal{F}_s$. It should be mentioned that there is only one stage codebook as a result of limited bits for quantization when the speech coding rate is 600 bit/s. As is shown in Tab. I, we will design a four-stage AbS vector quantizer to quantize the formant and pitch information for speech coding at 2400 bit/s, where each stage has 4096 codewords. As for speech coding at 1200 bit/s, a two-stage AbS vector quantizer is designed, where the first stage has 4096 codewords and the second stage has 2048 codewords. The bit allocation scheme of AbS SVQ is the same as which in [2]. Therefore, $\mathcal{F}_s$ will consist of Q^4, Q^2, Q codewords for speech coding at 2400, 1200 and 600 bit/s, respectively. If $Q = 1$, the AbS SVQ method will regress to be the conventional VQ method in [1]–[2].

TABLE I
BIT ALLOCATION OF ABS MULTISTAGE VQ.

Rate (bit/s)	Bits/ Frame	Energy (C_1)	Formant & Pitch ($C_2 - C_{60}$)
2400	54	6-bit SQ	(12-12-12-12)-bit AbS MSVQ
1200	27	4-bit SQ	(12-11)-bit AbS MSVQ

III. EXPERIMENTS AND RESULTS

A. Dataset and evaluation metrics

We use 4000 chosen utterances spoken by 40 speakers (20 males and 20 females) from CASIA Chinese database [13] to train the MFCCs codebook. The duration of training speech is ∼3.5 h, which contains 1,682,162 speech frames (30 ms time frames, 7.5 ms frame shifts). Another 96 chosen utterances from the same database, spoken by unseen 16 speakers (8 males and 8 females), are used as the test set. The duration of test speech is ∼5 m. Both the training speech and the test speech are downsampled at 8 kHz. We choose $N = 240, M = 60, p = 10$ to extract MFCCs vector and perform LP analysis.

We use three metrics to evaluate the quality of coded speech, which are frequency-weighted segmental SNR (fwsegSNRs) [14], perceptual evaluation of speech quality (PESQ) [15] and short-time objective intelligibility (STOI) [16]. FwsegSNRs and PESQ measures illustrate the overall speech quality while the STOI measure illustrates the speech intelligibility.

B. Objective evaluation of speech quality

The results of objective evaluation are shown in Tables II–IV, in which the reuslts of conventional VQ method are marked in underline while the best results are highlighted in bold. It is clearly illustrated that the proposed AbS VQ method yields substantially higher fwsegSNRs, PESQ, and STOI score than the conventional VQ method, which demonstrates that the speech quality is much better. Moreover, $\mathcal{F}_s$ will approach $\mathcal{F}$ with the increasing of Q, so the speech quality continues being improved. Specifically, the final improvement is impressive in the case of speech coding at 2400 bit/s, the average fwsegSNRs, PESQ and STOI score is improved by 1.2dB, 0.12 and 2%, respectively.

In addition, we can see that AbS MSVQ substantially yields better results than AbS SVQ, this is because it fully exploits the redundancy between the components of MFCCs vector. As mentioned above, larger Q will lead to more codewords in $\mathcal{F}_s$, hence, to make the trade-off between speech quality and computational complexity, we will choose $Q = 2, 4, 5$ in the case of speech coding at 2400, 1200 and 600 bit/s, respectively.

TABLE II
COMPARISON ON THE FWSEGSNRS (DB).

Rate (bit/s)	Quantization method	Q 1	2	3	4	5
2400	AbS SVQ	13.74	14.20	14.42	14.52	14.56
2400	AbS MSVQ	13.84	14.55	14.75	14.87	**14.94**
1200	AbS SVQ	11.91	12.25	12.44	12.51	12.56
1200	AbS MSVQ	12.05	12.41	12.56	12.65	**12.70**
600	AbS SVQ	10.75	10.91	10.97	11.01	**11.02**

TABLE III
COMPARISON ON THE STOI SCORE (%).

Rate (bit/s)	Quantization method	Q 1	2	3	4	5
2400	AbS SVQ	91.15	92.31	92.64	92.77	92.89
2400	AbS MSVQ	91.48	92.69	92.98	93.04	**93.35**
1200	AbS SVQ	88.24	89.52	89.83	89.98	90.02
1200	AbS MSVQ	88.32	89.58	89.86	90.12	**90.17**
600	AbS SVQ	84.98	85.86	85.89	86.00	**86.07**

TABLE IV
COMPARISON ON THE PESQ SCORE.

Rate (bit/s)	Quantization method	Q 1	2	3	4	5
2400	AbS SVQ	3.22	3.25	3.28	3.29	3.32
2400	AbS MSVQ	3.23	3.30	3.32	3.34	**3.35**
1200	AbS SVQ	2.91	2.98	2.99	3.01	3.02
1200	AbS MSVQ	2.94	3.00	3.01	3.02	**3.03**
600	AbS SVQ	2.63	2.66	2.66	2.67	**2.68**

C. Subjective listening test

14 native Chinese volunteers (7 males and 7 females) are invited to participate in a subjective listening test. Each volunteer is asked to rate the coded speech through the standard five point mean opinion score (MOS) [17]. Each volunteer is presented with two speech files (one male and one female) encoded by VQ and AbS MSVQ based MFCC codec at 2400,1200 and 600 bit/s, respectively. The results of subjective listening test are illustrated in Tab. V, which show a good match with PESQ evaluation.

TABLE V
SUBJECTIVE EVALUATION RESULTS.

Rate (bit/s)	2400	1200	600
AbS MSVQ	3.24±0.10	2.92±0.12	2.70±0.15
VQ [1]-[2]	3.16±0.15	2.82±0.14	2.65±0.15

D. Discussion on further performance improvement

Instead of utilizing the Moore-Penrose pseudo-inverse to reconstruct the speech power spectrum in (5), the synthesis stage of AbS VQ could be improved by some new algorithms [18]–[19], which is expected to further improve the speech quality. Furthermore, the speech waveforms could be reconstructed from the power spectrum using the improved algorithm in [20] rather than the inverse short-time Fourier transform magnitude algorithm in [21], which is also beneficial for yielding higher speech quality by taking into account the differences between voiced speech and unvoiced speech.

IV. CONCLUSION

In this paper, we propose a perceptually weighted AbS VQ approach for low bit rate MFCC codec. The objective of VQ is changed to minimize the perceptually weighted spectral reconstruction distortion rather than the distortion of MFCCs vector itself. A suboptimal codebook searching technique is proposed for practical implication. Objective and subjective tests show that the speech quality is substantially improved when compared to the output of current MFCC codec.

REFERENCES

[1] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Hybrid Scalar/Vector Quantization of Mel-Frequency Cepstral Coefficients for Low Bit-Rate Coding of Speech," *Data Compression Conference* (*DCC*), Mar. 2011, pp. 103–112.

[2] L. E. Boucheron, P. L. De Leon, and S. Sandoval, "Low Bit-Rate Speech Coding Through Quantization of Mel-Frequency Cepstral Coefficients," *IEEE Trans. Audio, Speech, and Language Process*. vol. 20, no. 2, pp. 610–619, Feb. 2012.

[3] ETSI ES 202 212, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm," 2005.

[4] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inforamtion theory*. vol. 44, no. 6, pp. 2325–2383, Oct. 1998.

[5] B. S. Atal and M. R. Schroeder, "Predictive coding of speech and subjective error criteria," *IEEE Trans. Acoust., Speech, Signal Process*. vol. ASSP-27, pp. 247–254, Mar. 1979.

[6] P. Kroon and B. S. Atal, "Predictive coding of speech using analysis-by-synthesis techniques," in *Advances in Speech Signal Process*. S. Furui and M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 141–164.

[7] P. Kroon and E. Deprettere, "A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbit/s," *IEEE Journal Select. Areas Commun.* vol. 6, no. 2, pp. 353–363, Feb. 1988.

[8] E. George and M. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech and Audio Process*. vol. 5, no. 5, pp. 389–406, Sep. 1997.

[9] S. Chatterjee and T.V. Sreenivas, "Analysis-by-Synthesis based switched transform domain split VQ using Gaussian mixture model," *IEEE Int. Conf. Acoust., Speech, Signal Process.* (*ICASSP*), May 2009, pp. 4117–4120.

[10] A. Shirazinia, S. Chatterjee, M. Skoglund, "Analysis-by-Synthesis Quantization for Compressed Sensing Measurements," *IEEE Trans. Signal Process*. vol. 61, no. 22, pp. 5789–5800, Nov. 2013.

[11] J. H. Chen, R. V. Cox, Y. C. Lin, et al, "A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard," *IEEE Journal Select. Areas Commun.* vol. 10, no. 5, pp. 830–849, Jun. 1992.

[12] 3GPP2 C.S0014-D, "Enhanced Variable Rate Codec, Speech Service Options 3, 68, 70 and 73 for Wideband Spread Spectrum Digital Systems," Oct. 2010.

[13] 2015 [Online]. Available: http://www.chineseldc.org.

[14] J. Tribolet, P. Noll, B. McDermott, etc, "A study of complexity and quality of speech waveform coders," *IEEE Int. Conf. Acoust., Speech, Signal Process.* (*ICASSP*), May 1978, pp. 586–590.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, etc, "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs," *IEEE Int. Conf. Acoust., Speech, Signal Process.* (*ICASSP*), May 2001 vol. II, pp. 749–752.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, etc, "An Algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Acoust., Speech, and Signal Process*. vol. 19, no. 7. pp. 2125–2136, Jul. 2011.

[17] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[18] G. Min, X. W. Zhang, J. B. Yang, etc, "Speech reconstruction from mel-frequency cepstral coefficients via ℓ_1-norm minimization," *IEEE Int. Workshop on Multimedia Signal Process.* (*MMSP'15*), Oct. 2015, pp. 1–5.

[19] G. Min, X. W. Zhang, J. B. Yang, etc, "Speech Reconstruction from MFCC based on nonnegative and sparse priors," *IEICE Trans. Fundamentals.* vol. E98A, no. 7, pp. 1540–1543, Jul. 2015.

[20] W. B. Jiang, R. D. Ying, P. L. Liu, "Speech reconstruction for MFCC-based low bit-rate speech coding," *IEEE Int. Conf. on Multimedia & Expro* (*ICME*), Oct. 2014, pp. 1–6.

[21] D. W. Griffin and J. S. Lim, "Signal estimation from modified short time fourier transform," *IEEE Trans. Acoust., Speech, and Signal Process*. vol. 32, no. 2, pp. 236–243, Apr. 1984.